visible as it is in brighter areas. Another opportunity for bit savings is presented in textured areas of the picture where high frequency coefficient error is much less visible than in relatively flat areas. Brightness and texture weighting require analysis of the original image since these areas may be well predicted. Finally, distortion is easily masked by limiting its duration to one or two frames. This effect is most profitably used after scene changes where the first frame or two can be greatly distorted without perceptible distortion at normal speed.

When quantizing transform coefficients, the differing perceptual importance of the various coefficients can be exploited by "allocating the bits" to shape the quantization noise into the perceptually less important areas. This can be accomplished by varying the relative step-sizes of the quantizers for the different coefficients. The perceptually important coefficients may be quantized with a finer step size than the others. For example, low spatial frequency coefficients may be quantized finely, while the less important high frequency coefficients may be quantized more coarsely. A simple method to achieve different step-sizes is to normalize or weight each coefficient based on its visual importance. All of the normalized coefficients may then be quantized in the same manner, such as rounding to the nearest integer (uniform quantization). Normalization or weighting effectively scales the quantizer from one coefficient to another. The video compression system utilizes perceptual weighting, where the different DCT coefficients are weighted according to a perceptual criterion prior to uniform quantization. The perceptual weighting is determined by quantizer matrices. The ATV video compression system allows for modifying the quantizer matrices before each picture.

5.8.3 Adaptive intra-quantizer and non-intra-quantizer matrices

The video coding syntax allows the quantizer matrices to be specified for every picture for improved coding efficiency. A certain probability distribution is associated with the variable-length codes (VLC) for quantized coefficients. Although one cannot change the VLC distribution to match the actual distribution of the data, the quantizer matrices can be adjusted to help match the distribution of the data to the distribution of the VLC. Over the course of encoding the frame data, the variance of each spatial frequency band may be calculated for both intra-data and non-intra-data. One method for choosing quantizer matrices involves applying upper and lower bounds per band to ensure reasonable operation in all cases.

Transmitting the quantizer matrices costs bits in the compressed data stream. If sent with every picture in the 60 fps progressive mode, the matrices consume 0.32% of the channel bandwidth. This modest amount of overhead can be reduced by updating the quantizer matrix less frequently, or only when the difference between the desired quantizer matrix and the prevailing quantizer matrix becomes significant.

Sufficient compression cannot be achieved unless a large fraction of the DCT coefficients are dropped and therefore not selected for transmission. The coefficients which are not selected are assumed to have zero value in the decoder.

The DC coefficients are coded differently to take advantage of high spatial correlation. For example, when intra-coded, the first DC coefficient in a slice is sent absolutely; the following DC coefficients are sent as differences.

5.9 Entropy coding of video data

Quantization creates an efficient discrete representation for the data to be transmitted. Codeword assignment takes the quantized values and produces a digital bit stream for transmission. Hypothetically, the quantized values could be simply represented using uniform or fixed-length codewords. Under this approach, every quantized value would be represented with the same number of bits. Greater efficiency, in terms of bit rate, can be achieved by employing entropy coding. Entropy coding attempts to exploit the statistical properties of the signal to be encoded. A signal, whether it is a pixel value or a transform coefficient, has a certain amount of information, or entropy, based on the probability of the different possible values or events occurring. For example, an event that occurs infrequently conveys much more new information than one that occurs often. By realizing that some events occur more frequently than others, the average bit rate may be reduced.

5.9.1 Huffman coding

Huffman coding, which is utilized in the video compression system, is one of the most common entropy coding schemes. In Huffman coding, a code book is generated which can approach the minimum average description length (in bits) of events, given the probability distribution of all the events. Events which are more likely to occur will be assigned shorter length codewords while those which are less likely to occur will be assigned longer length codewords.

5.9.2 Run-length coding

In video compression, most of the transform coefficients are frequently quantized to zero. There may be a few non-zero low-frequency coefficients and a sparse scattering of non-zero high-frequency coefficients, but the great majority of coefficients may have been quantized to zero. To exploit this phenomenon the two-dimensional array of transform coefficients is reformatted and prioritized into a one-dimensional sequence through either a zigzag or alternate scanning process. This results in most of the important non-zero coefficients (in terms of energy and visual perception) being grouped together early in the sequence. They will be followed by long runs of coefficients that are quantized to zero. These zero-valued coefficients can be efficiently represented through run-length encoding. In run-length encoding, the number (run) of consecutive zero coefficients before a non-zero coefficient is encoded, followed by the non-zero coefficient value. The run-length and the coefficient value can be entropy coded, either separately or jointly. The scanning separates most of the zero and the non-zero coefficients into groups, thereby enhancing the efficiency of the run-length encoding process. Also, a special end-of-block (EOB) marker is used to signify when all of the remaining coefficients in the sequence are equal to zero. This approach can be extremely efficient, yielding a significant degree of compression.

5.9.3 Zigzag scan and alternate scan

As indicated above, the array of 64 DCT coefficients is arranged in a one-dimensional vector before run-length/amplitude code-word assignment. Two different one-dimensional arrangements, or *scan types*, are allowed, which are generally referred to as zigzag scan (shown in Figure 5.6a) and alternate scan (shown in Figure 5.6b). The scan type is specified before coding each picture, and is permitted to vary from picture-to-picture.

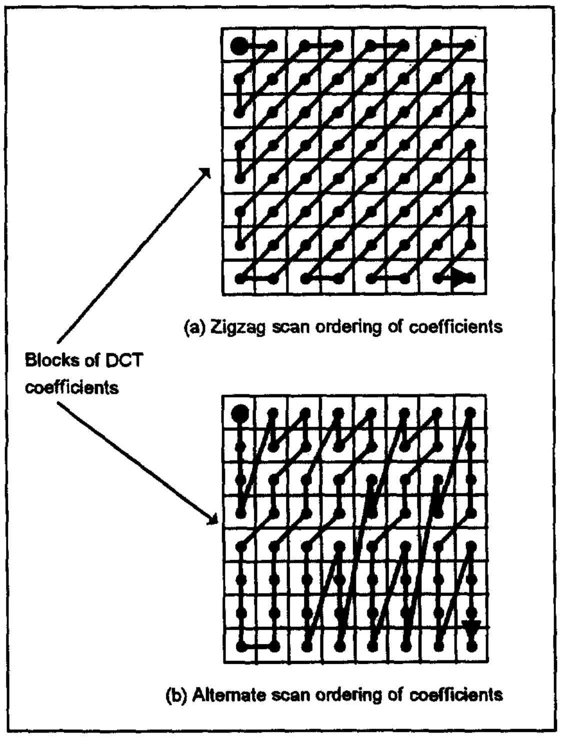
(a) Zigzag scan ordering of coefficients
Blocks of DCT coefficients
(b) Alternate scan ordering of coefficients

Figure 5.6. Scanning of coefficient blocks.

5.10 Channel buffer

Whenever entropy coding is employed, the bit rate produced by the encoder is variable and is a function of the video statistics. Since the bit rate permitted by the transmission system is less than the peak bit rate produced by the variable-length coder, a *channel buffer* is necessary at the decoder. The buffering must be carefully designed. What is needed is some form of buffer control that would allow efficient allocation of bits to encode the video while ensuring that no overflow or underflow occurs.

The buffer control typically involves a feedback mechanism to the compression algorithm whereby the amplitude resolution (quantization) and/or spatial, temporal and color resolution may be varied in accordance with the instantaneous bit rate requirements. If the bit rate decreases significantly, a finer quantization can be performed to increase it.

As indicated above the peak bit rate produced by variable-length coding will typically fluctuate dramatically, and will frequently exceed the desired peak channel bandwidth allocation.[8] This situation is handled by the presence of a channel buffer at each decoder for temporary storage of the coded bit stream. The Digital Television Standard specifies a channel buffer size of 8 Mbits.

A model buffer is defined in the video coding system as a reference for manufacturers of both encoders and decoders to ensure interoperability. An encoder will control its production of bits so that the model buffer does not overflow or underflow.[9]

In order to avoid overflow or underflow of the model buffer, an encoder may maintain measures of buffer occupancy and scene complexity. When the encoder needs to reduce the number of bits produced, it may do so by increasing the general value of the quantizer scale, which will increase picture degradation. When it is able to produce more bits, it may decrease the quantizer scale thus decreasing picture degradation.

5.11 Interface to system multiplexer

The bit stream produced by the video encoder is passed to the transport encoding system for multiplexing with audio and ancillary data, "lip-synch", and scheduling for delivery.

5.12 Decoder block diagram

As shown in Figure 5.7 the video decoder contains elements which invert, or undo, the processing performed in the encoder. The incoming coded video bit stream is placed in the channel buffer. Bits are removed from the channel buffer by a variable-length decoder (VLD).

The VLD reconstructs 8-by-8 arrays of quantized DCT coefficients by decoding run-length/amplitude codes and appropriately distributing the coefficients according to the scan type used. These coefficients are de-quantized and transformed by the inverse discrete cosine transform (IDCT) to obtain pixel values or prediction errors.

In the case of interframe prediction the decoder uses the received motion vectors to perform the same prediction operation as was done in the encoder. The prediction errors are summed with the results of motion compensated prediction to produce pixel values.

5.12.1 Error concealment capability

When transmission errors occur, a decoder may act to minimize the perceived picture degradation. This process is discussed in more detail in Section 10.2.6.

[8] Note that the average bit rate, by definition, cannot be permitted to exceed the peak allocated channel bandwidth.

[9] Buffer underflow is actually permitted in the case of low-delay bit streams which do not contain B-frames. In such cases underflow might occur due to an unusually difficult picture which requires a particularly large number of bits. The result of underflow is the repeat of one or more pictures, which are presented in lieu of pictures that were skipped in the encoder.

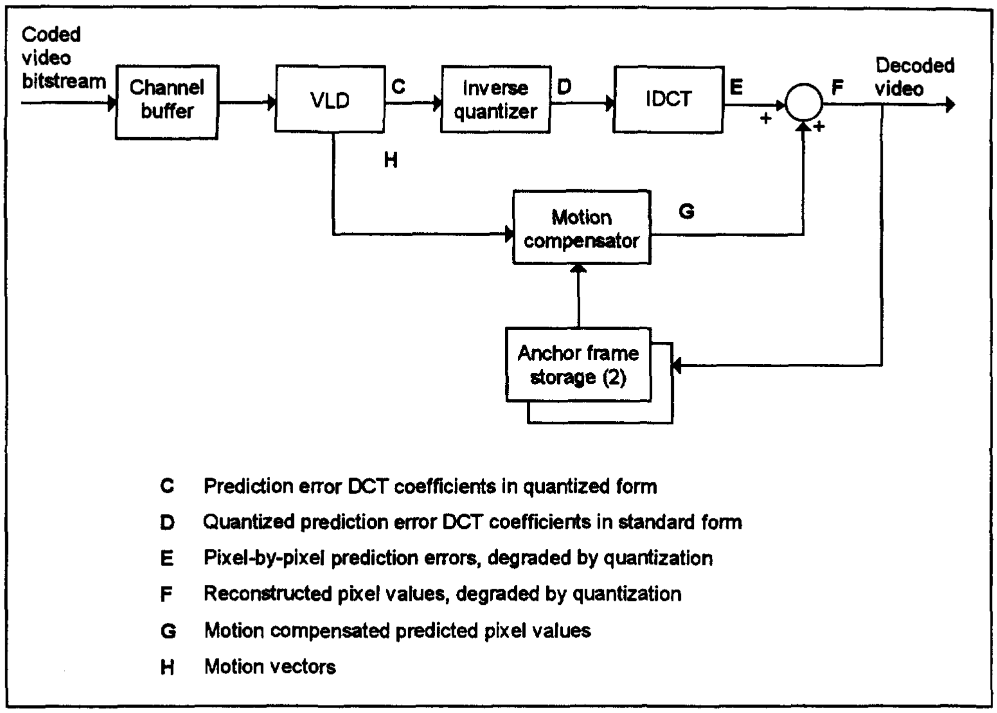
Coded video bitstream
Channel buffer
VLD
C
Inverse quantizer
D
IDCT
E
+
+
F
Decoded video
H
Motion compensator
G
Anchor frame storage (2)
C Prediction error DCT coefficients in quantized form
D Quantized prediction error DCT coefficients in standard form
E Pixel-by-pixel prediction errors, degraded by quantization
F Reconstructed pixel values, degraded by quantization
G Motion compensated predicted pixel values
H Motion vectors

Figure 5.7. Decoder block diagram.

5.12.2 Frame store for decoded pictures

As described above, pixel values are decoded from the incoming bit stream. In the case of decoded anchor frames (I or P-frames) these values must be stored in a frame buffer for subsequent use as prediction references. When B-frames are used the anchor frame storage also allows for the necessary frame re-ordering for display.

5.13 Concatenated sequences

The MPEG-2 standard which underlies the Digital Television Standard clearly specifies the behavior of a compliant video decoder when processing a single video sequence. A coded video sequence commences with a sequence header, may contain some repeated sequence headers and one or more coded pictures, and is terminated by an end-of-sequence code. A number of parameters are specified in the sequence header that are required to remain constant throughout the duration of the sequence. The sequence level parameters include, but are not limited to:

- Horizontal and vertical resolution
- Frame rate
- Aspect ratio
- Chroma format
- Profile and level
- All-progressive indicator

- Video buffering verifier (VBV) size
- Maximum bit rate

It is envisioned that it will be common for coded bit streams to be spliced for editing, insertion of commercial advertisements, and other purposes in the video production and distribution chain. If one or more of the sequence level parameters differ between the two bit streams to be spliced, then an end-of-sequence code must be inserted to terminate the first bit stream and a new sequence header must exist at the start of the second bit stream. Thus the situation of concatenated video sequences arises.

While the MPEG-2 standard specifies the behavior of video decoders when processing a single sequence, it does not place any requirements on the handling of concatenated sequences. Specification of the decoding behavior in the former case is feasible because the MPEG-2 standard places constraints on the construction and coding of individual sequences. These constraints prohibit channel buffer overflow and coding the same field parity for two consecutive fields. The MPEG-2 standard does not prohibit these situations at the junction between two coded sequences and it likewise does not specify the behavior of decoders in this case.

While it is recommended, the Digital Television Standard does not require the production of well-constrained concatenated sequences. Well-constrained concatenated sequences are defined as having the following characteristics:

- The extended decoder buffer never overflows, and may only underflow in the case of low-delay bit streams. Here "extended decoder buffer" refers to the natural extension of the MPEG-2 decoder buffer model to the case of continuous decoding of concatenated sequences.
- When field parity is specified in two coded sequences which are concatenated, the parity of the first field in the second sequence is opposite that of the last field in the first sequence.
- Whenever a progressive sequence is inserted between two interlaced sequences, the exact number of progressive frames shall be such that the parity of the interlaced sequences is preserved as if no concatenation had occurred.

5.14 Guidelines for refreshing

While the Digital Television Standard does not require refreshing at less than the intra-macroblock refresh rate as defined in IEC/ISO 13818-2, the following is recommended:

- In a system which uses periodic transmission of I-frames for refreshing, the frequency of occurrence of I-frames will determine the channel-change time performance of the system. In this case, it is recommended that I-frames be sent at least once every 0.5 second in order to have acceptable channel-change performance. It is recommeded also that sequence layer information be sent before every I-frame.

- In order to spatially localize errors due to transmission, intra-coded slices should contain fewer macroblocks than the maximum number allowed by the Standard. It is recommended that there be four to eight slices in a horizontal row of intra-coded macroblocks for the intra-coded slices in the I-frame refresh case as well as for the intraframe coded regions in the progressive refresh case. The size of non-intra-coded slices can be larger than that of intra-coded slices.

6. AUDIO SYSTEMS

This section describes the audio coding technology and gives guidelines as to its use. Information of interest to both broadcasters (and other program providers) and receiver manufacturers is included. The audio system is fully specified in Annex B of the Digital Television Standard and is based on the Digital Audio Compression (AC-3) Standard, with some limitations on bit rate, sampling rate, and audio coding mode.

6.1 Audio system overview

As illustrated in Figure 6.1, the audio subsystem comprises the audio encoding/decoding function and resides between the audio inputs/outputs and the transport subsystem. The audio encoder(s) is (are) responsible for generating the audio elementary stream(s) which are encoded representations of the baseband audio input signals. The flexibility of the transport system allows multiple audio elementary streams to be delivered to the receiver. At the receiver, the transport subsystem is responsible for selecting which audio streams(s) to deliver to the audio subsystem. The audio subsystem is responsible for decoding the audio elementary stream(s) back into baseband audio.

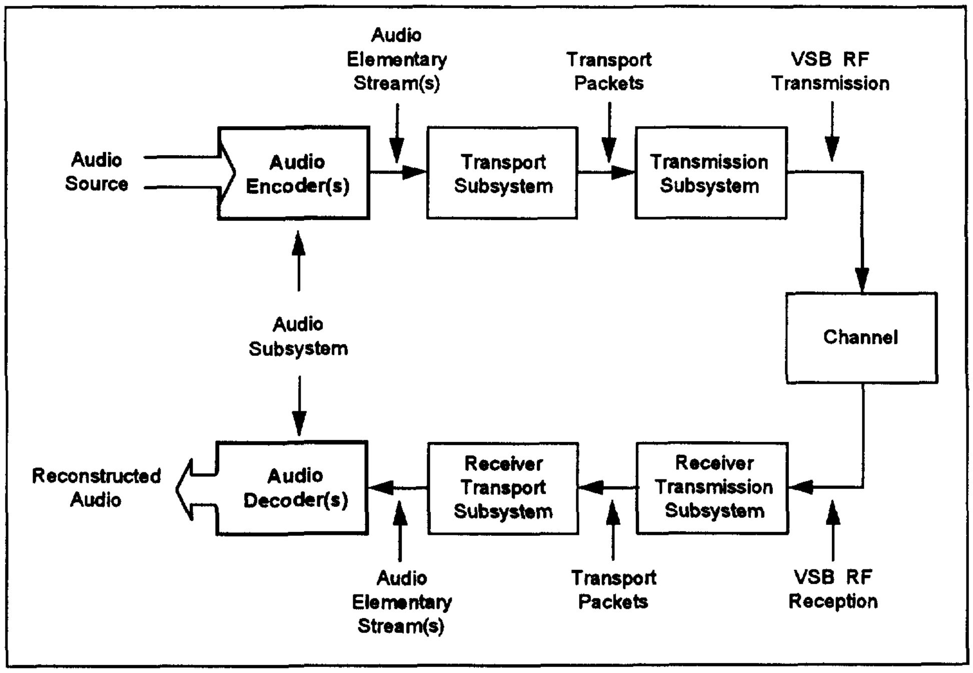

Figure 6.1. Audio subsystem within the digital television system.

An audio program source is encoded by a digital television audio encoder. The output of the audio encoder is a string of bits that represent the audio source, and is referred to as an *audio elementary stream*. The transport subsystem packetizes the audio data into PES packets which are then further packetized into transport packets. The transmission subsystem converts the transport packets into a modulated RF signal for transmission to the receiver. At the receiver, the received signal is demodulated by the receiver transmission subsystem. The receiver transport subsystem converts the received audio packets back into an audio elementary stream which is decoded by the digital television audio decoder. The partitioning shown is conceptual, and practical implementations may differ. For example, the transport processing may be broken into two blocks; one to perform PES packetization, and the second to perform transport packetization. Or, some of the transport functionality may be included in either the audio coder or the transmission subsystem.

6.2 *Audio encoder interface*

The audio system accepts baseband audio inputs with up to six audio channels per audio program bit stream. The channelization is consistent with ITU-R Recommendation BS-775, "*Multi-channel stereophonic sound system with and without accompanying picture*". The six audio channels are: Left, Center, Right, Left Surround, Right Surround, and Low Frequency Enhancement (LFE). Multiple audio elementary bit streams may be conveyed by the transport system.

The bandwidth of the LFE channel is limited to 120 Hz. The bandwidth of the other (main) channels is limited to 20 kHz. Low frequency response may extend to DC, but is more typically limited to approximately 3 Hz (-3 dB) by a DC blocking high-pass filter. Audio coding efficiency (and thus audio quality) is improved by removing DC offset from audio signals before they are encoded.

There are two classes of audio programs defined: main service, and associated service. Main services contain all of the audio program content, with the possible exception of dialogue. Associated services are intended to be used along with a main service, and can contain dialogue, commentary, descriptive video, dialogue intended for the hearing impaired, voice-overs, or emergency messages.

6.2.1 Input source signal specification

Audio signals which are input to the audio system may be in analog or digital form.

6.2.1.1 High-pass filtering

Audio signals should have any DC offset removed before being encoded. If the audio encoder does not include a DC blocking high-pass filter, the audio signals should be high-pass filtered before being applied to the audio encoder.

6.2.1.2 Analog input

For analog input signals, the input connector and signal level are not specified. Conventional broadcast practice may be followed. One commonly used input connector is the 3-pin XLR female (the incoming audio cable uses the male connector) with pin 1 ground, pin 2 hot or positive, and pin 3 neutral or negative.

6.2.1.3 Digital input

For digital input signals, the input connector and signal format are not specified. Commonly used formats such as the AES 3-1992 two channel interface may be used. When multiple two channel inputs are used, the preferred channel assignment is:

Pair 1:	Left, Right
Pair 2:	Center, LFE
Pair 3:	Left Surround, Right Surround

6.2.1.4 Sampling frequency

The system conveys digital audio sampled at a frequency of 48 kHz, locked to the 27 MHz system clock. If analog signal inputs are employed, the A/D converters should sample at 48 kHz. If digital inputs are employed, the input sampling rate shall be 48 kHz, or the audio encoder shall contain sampling rate converters which convert the sampling rate to 48 kHz. The sampling rate at the input to the audio encoder must be locked to the video clock for proper operation of the audio subsystem.

6.2.1.5 Resolution

In general, input signals should be quantized to at least 16-bit resolution. The audio compression system can convey audio signals with up to 24-bit resolution.

6.2.2 Output signal specification

Conceptually, the output of the audio encoder is an elementary stream which is formed into PES packets within the transport subsystem. It is possible that digital television systems will be implemented wherein the formation of audio PES packets takes place within the audio encoder. In this case, the output(s) of the audio encoder(s) would be PES packets. Physical interfaces for these outputs (elementary streams and/or PES packets) may be defined as voluntary industry standards by SMPTE or other standards organizations.

6.3 AC-3 digital audio compression

6.3.1 Overview and basics of audio compression

The audio compression system conforms with the Digital Audio Compression (AC-3) Standard specified in ATSC Doc. A/52. The audio compression system is considered a constrained subset of that Standard. The constraints are specified in Annex B of the Digital Television Standard. By conforming with the standardized syntax in ATSC Doc. A/52, the system employs an audio compression system which is interoperable across many different media, and is appropriate for use in a multitude of applications.

A major objective of audio compression is to represent an audio source with as few bits as possible, while preserving the level of quality required for the given application. Audio compression has two major applications. One is efficient utilization of channel bandwidth for video transmission systems. The other is reduction of storage requirements. Both of these applications apply to the digital television system.

The audio compression system consists of three basic operations, as shown in Figure 6.2. In the first stage, the representation of the audio signal is changed from the time domain to the frequency domain, which is a more efficient domain in which to perform psychoacousticly based audio compression. The resulting frequency domain coefficients are what are then encoded. The frequency domain coefficients may be coarsely quantized because the resulting quantizing noise will be at the same frequency as the audio signal, and relatively low signal to noise ratios are acceptable due to the phenomena of psychoacoustic masking. The bit allocation operation determines, based on a psychoacoustic model of human hearing, what actual SNR is acceptable for each individual frequency coefficient. Finally, the frequency coefficients are coarsely quantized to the necessary precision and formatted into the audio elementary stream. The basic unit of encoded audio is the AC-3 sync frame, which represents 1536 audio samples. Each sync frame of audio is a completely independent encoded entity. The elementary bit stream contains the information necessary to allow the audio decoder to perform the identical (to the encoder) bit allocation. This allows the decoder to unpack and de-quantize the elementary bit stream frequency coefficients, resulting in the reconstructed frequency coefficients. The synthesis filterbank is the inverse of the analysis filterbank, and converts the reconstructed frequency coefficients back into a time domain signal.

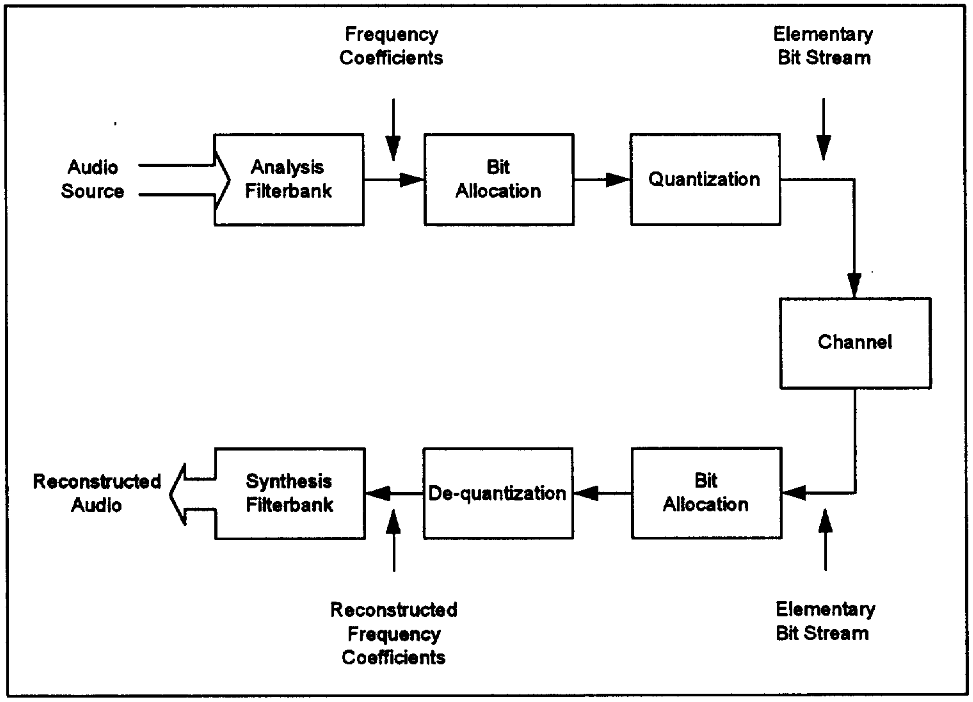

Figure 6.2. Overview of audio compression system.

6.3.2 Transform filter bank

The process of converting the audio from the time domain to the frequency domain requires that the audio be blocked into overlapping blocks of 512 samples. For every 256 new audio samples, a 512 sample block is formed from the 256 new samples, and the 256 previous samples. Each audio sample is represented in two audio blocks, and thus the number of samples to be processed initially is doubled. The overlapping of blocks is necessary in order to prevent audible blocking artifacts. New audio blocks are formed every 5.33 ms. A group of 6 blocks are coded into one AC-3 sync frame.

6.3.2.1 Window function

Prior to being transformed into the frequency domain, the block of 512 time samples is windowed. The windowing operation involves a vector multiplication of the 512 point block with a 512 point window function. The window function has a value of 1.0 in its center, and tapers down to almost zero at its ends. The shape of the window function is such that the overlap/add processing at the decoder will result in a reconstruction free of blocking artifacts. The window function shape also determines the shape of each individual filterbank filter.

6.3.2.2 Time division aliasing cancellation transform

The analysis filterbank is based on the fast Fourier transform. The particular transformation employed is the oddly stacked time domain aliasing cancellation (TDAC) transform. This particular transformation is advantageous because it allows the 100% redundancy which was introduced in the blocking process to be removed. The input to the TDAC transform is 512 windowed time domain points, and the output is 256 frequency domain coefficients.

6.3.2.3 Transient handling

When extreme time domain transients exist (such as an impulse or a castanet click), there is a possibility that quantization error, incurred in coarsely quantizing the frequency coefficients of the transient, will become audible due to time smearing. The quantization error within a coded audio block is reproduced throughout the block. It is possible for the portion of the quantization error which is reproduced prior to the impulse to be audible. Time smearing of quantization noise may be reduced by altering the length of the transform which is performed. Instead of a single 512 point transform, a pair of 256 point transforms may be performed, one on the first 256 windowed samples, and one on the last 256 windowed samples. A transient detector in the encoder determines when to alter the transform length. The reduction in transform length prevents quantization error from spreading more than a few milliseconds in time, which is adequate to prevent its audibility.

6.3.3 Coded audio representation

The frequency coefficients which result from the transformation are converted to a binary floating point notation. The scaling of the transform is such that all values are smaller than 1.0. An example value in binary notation (base 2) with 16-bit precision would be:

$$0.0000\ 0000\ 1010\ 1100_2$$

The number of leading zeroes in the coefficient, 8 in this example, becomes the raw exponent. The value is left shifted by the exponent, and the value to the right of the decimal point (1010 1100) becomes the normalized mantissa to be coarsely quantized. The exponents and the coarsely quantized mantissas are encoded into the bit stream.

6.3.3.1 Exponent coding

Some processing is applied to the raw exponents in order to reduce the amount of data required to encode them. First, the raw exponents of the 6 blocks to be included in a single AC-3 sync frame are examined for block-to-block differences. If the differences are small, a single exponent set is generated which is useable by all 6 blocks, thus reducing the amount of data to be encoded by a factor of 6. If the exponents undergo significant changes within the frame, then exponent sets are formed over blocks where the changes are not significant. Due to the frequency response of the individual filters in the analysis filter bank, exponents for adjacent frequencies rarely differ by more than ±2. To take advantage of this fact, exponents are encoded differentially in frequency. The first exponent is encoded as an absolute, and the difference between the current exponent and the following exponent is then encoded. This reduces the exponent data rate by a factor of 2. Finally, where the spectrum is relatively flat, or an exponent set only covers 1-2 blocks, differential exponents may be shared across 2 or 4 frequency coefficients, for an additional savings of a factor of 2 or 4.

The final coding efficiency for exponents is typically 0.39 bits/exponent (or 0.39 bits/sample since there is an exponent for each audio sample). Exponents are only coded up to the frequency needed for the perception of full frequency response. Typically, the highest audio frequency component in the signal which is audible is at a frequency lower than 20 kHz. In the case that signal components above 15 kHz are inaudible, only the first 75% of the exponent values are encoded, reducing the exponent data rate to <0.3 bits/sample.

The exponent processing changes the exponent values from their original values. The encoder generates a local representation of the exponents which is identical to the decoded representation which will be used by the decoder. The decoded representation is then used to shift the original frequency coefficients to generate the normalized mantissas which are quantized.

6.3.3.2 Mantissas

The frequency coefficients produced by the analysis filterbank have useful precision dependent on the wordlength of the input PCM audio samples, and the precision of the transform computation. Typically this precision is on the order of 16-18 bits, but may be as high as 24 bits. Each normalized mantissa is quantized to a precision between 0 and 16 bits. The goal of audio compression is to maximize the audio quality at a given bit rate. This requires an optimum (or near optimum) allocation of the available bits to the individual mantissas.

6.3.4 Bit allocation

The number of bits allocated to each individual mantissa value is determined by the bit allocation routine. The identical core routine is run in both the encoder and the decoder, so that each generates the identical bit allocation.

6.3.4.1 Backward adaptive

The core bit allocation algorithm is considered backwards adaptive, in that the some of the encoded audio information within the bit stream (fed back into the encoder) is used to compute the final bit allocation. The primary input to the core allocation routine is the decoded exponent values, which give a general picture of the signal spectrum. From this version of the signal spectrum, a masking curve is calculated. The calculation of the masking model is based on a model of the human auditory system. The masking curve indicates, as a function of frequency, the level of quantizing error which may be tolerated. Subtraction (in the log power domain) of the masking curve from the signal spectrum yields the required SNR as a function of frequency. The required SNR values are mapped into a set of bit allocation pointers (baps) which indicate which quantizer to apply to each mantissa.

6.3.4.2 Forward adaptive

The AC-3 encoder may employ a more sophisticated psychoacoustic model than that used by the decoder. The core allocation routine used by both the encoder and the decoder makes use of a number of adjustable parameters. If the encoder employs a more sophisticated psychoacoustic model than that of the core routine, the encoder may adjust these parameters so that the core routine produces a better result. The parameters are inserted into the bit stream by the encoder and fed forward to the decoder.

In the event that the available bit allocation parameters do not allow the ideal allocation to be generated, the encoder can insert explicit codes into the bit stream to alter the computed masking curve, and thus the final bit allocation. The inserted codes indicate changes to the base allocation, and are referred to as delta bit allocation codes.

6.3.5 Rematrixing

When the AC-3 coder is operating a two channel stereo mode, an additional processing step is inserted in order to enhance interoperability with Dolby Surround 4-2-4 matrix encoded programs. The extra step is referred to as *rematrixing*.

The signal spectrum is broken into four distinct rematrixing frequency bands. Within each band, the energy of the Left, Right, Sum, and Difference signals are determined. If the largest signal energy is in the Left or Right channels, the band is encoded normally. If the dominant signal energy is in the Sum or Difference channel, then those channels are encoded instead of the Left and Right channels. The decision as to whether to encode Left and Right, or Sum and Difference is made on a band-by-band basis and is signaled to the decoder in the encoded bit stream.

6.3.6 Coupling

In the event that the number of bits required to encode the audio signals transparently exceeds the number of bits which are available, the encoder may invoke coupling. Coupling involves combining the high frequency content of individual channels and sending the individual channel signal envelopes along with the combined coupling channel. The psychoacoustic basis for coupling is that within narrow frequency bands the human ear detects high frequency localization based on the signal envelope rather than the detailed signal waveform.

The frequency above which coupling is invoked, and the channels which participate in the process, are determined by the AC-3 encoder. The encoder also determines the frequency banding structure used by the coupling process. For each coupled channel and each coupling band, the encoder creates a sequence of coupling coordinates. The coupling coordinates for a particular channel indicate what fraction of the common coupling channel should be reproduced out of that particular channel output. The coupling coordinates represent the individual signal envelopes for the channels. The encoder determines the frequency with which coupling coordinates are transmitted. When coupling is in use, coupling coordinates are always sent in block 0 of a frame. If the signal envelope is steady, the coupling coordinates do not need to be sent every block, but can be reused by the decoder until new coordinates are sent. The encoder determines how often to send new coordinates, and can send them as often as every block (every 5.3 ms).

6.4 Bit stream syntax

6.4.1 Sync frame

The audio bit stream consists of a repetition of audio frames which are referred to as AC-3 sync frames. Shown in Figure 6.3, each AC-3 sync frame is a self contained entity consisting of synchronization information (SI), bit stream information (BSI), 32 ms of encoded audio, and a CRC error check code. Every sync frame is the same size (number of bits) and contains six encoded audio blocks. The sync frame may be considered an audio access unit. Within SI is a 16-bit sync word, an indication of audio sample rate (48 kHz for the digital television system), and an indication of the size of the audio frame (which indicates bit rate).

6.4.2 Splicing, insertion

The ideal place to splice encoded audio bit streams is at the boundary of a sync frame. If a bit stream splice is performed at the sync frame boundary, the audio decoding will proceed without interruption. If a bit stream splice is performed randomly, there will be an audio interruption. The frame which is incomplete will not pass the decoder's error detection test and this will cause the decoder to mute. The decoder will not find sync in its proper place in the next frame, and will enter a sync search mode. Once the sync code of the new bit stream is found, synchronization will be achieved, and audio reproduction may begin once again. The outage will be on the order of two frames, or about 64 ms. Due to the windowing process of the filterbank, when the audio goes to mute there will be a gentle fade down over a period of 2.6 ms. When the audio is recovered, it will fade up over a period of 2.6 ms. Except for the approximately 64 ms of time during which the audio is muted, the effect of a random splice of an AC-3 elementary stream is relatively benign.

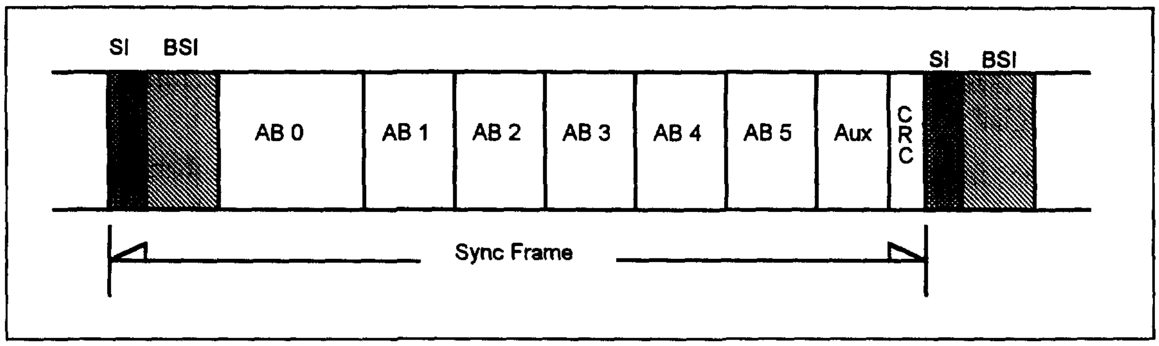

Figure 6.3. AC-3 synchronization frame.

6.4.3 Error detection codes

Each AC-3 sync frame ends with a 16-bit CRC error check code. The decoder may use this code to determine whether a frame of audio has been damaged or is incomplete. Additionally, the decoder may make use of error flags provided by the transport system. In the case of detected errors, the decoder may try to perform error concealment, or may simply mute.

6.5 Loudness and dynamic range

6.5.1 Loudness normalization

It is important for the digital television system to provide uniform subjective loudness for all audio programs. Consumers find it very annoying when audio levels fluctuate between broadcast channels (observed when channel hopping), or between program segments on a particular channel (commercials much louder than the entertainment). One element which is found in most audio programming is the human voice. Achieving an approximate level match for dialogue (spoken in a normal voice, not shouting or whispering) amongst all audio programming is a desirable goal. The AC-3 audio system provides syntactical elements which make this goal achievable.

There is (currently) no regulatory limit as to how loud dialogue may be in an encoded bit stream. Since the digital audio coding system can provide more than 100 dB of dynamic range, there is no technical reason for dialogue to be encoded anywhere near 100% as is commonly done in NTSC television. However, there is no assurance that all program channels, or all programs or program segments on a given channel, will have dialogue encoded at the same (or even similar) level. Lacking a uniform coding level for dialogue (which would imply a uniform headroom available for all programs) there would be inevitable audio level fluctuations between program channels or even between program segments.

Encoded AC-3 elementary bit streams are tagged with an indication (dialnorm) of the subjective level at which dialogue has been encoded. Different audio programs may be encoded with differing amounts of headroom above the level of dialogue in order to allow for dynamic music and sound effects. The digital television receiver (and all AC-3 decoders) are able to use the value of dialnorm to adjust the reproduced level of audio programs so that different received programs have their spoken dialogue reproduced at a uniform level. Some receiver designs may even offer the listener an audio volume control calibrated in absolute sound pressure level. The listener could dial up the desired SPL for dialogue, and the receiver would scale the level of every decoded audio program so that the dialogue is always reproduced at the desired level.

The BSI portion of the sync frame contains the 5-bit dialnorm field which indicates the level of average spoken dialogue within the encoded audio program. The indication is relative to the level of a full scale 1 kHz sinewave. The measurement of dialogue level is done by a method which gives a subjectively accurate value. The measurement of subjective loudness is not an exact science, and new measurement techniques will be developed in the future. A measurement method which is currently available and quite useful is the "A" weighted integrated measurement (L_{Aeq}). This measurement method should be used until a more accurate method is standardized and available in practical equipment. Any new measurement methodology which is developed should be normalized (scaled) so that its results generally match those of the L_{Aeq} method.

It is important for broadcasters and others who deliver encoded audio bit streams to ensure that the value of dialnorm is correct. Incorrect values will lead to unwelcome level fluctuations in consumer homes. The worst case example of incorrect (or abusive) setting of dialnorm would be to broadcast a commercial message which indicates dialogue at a low level, but which is actually encoded with dialogue at full level. This would result in the commercial message being reproduced at the same level as a full scale explosion in a feature film (>100 dB SPL in some home theatre setups!). If such abuses occur, there may be a demand for regulatory enforcement of audio levels. Fortunately, bit streams which contain an incorrect value of dialnorm are easily corrected by simply changing the value of the 5-bit dialnorm field in the BSI header.

There are two primary methods which broadcast organizations may employ to ensure that the value of dialnorm is set correctly. The first method is to select a suitable dialogue level for use with all programming and conform all baseband audio programs to this level prior to AC-3 encoding. Then the value of dialnorm can be set to one common value for all programs which are encoded. Conforming all programs to a common dialogue level may mean that for some programs the audio level never approaches 100% digital level (since they have to be reduced in gain), while for other programs non-reversible (by the receiver) limiting must be engaged in order to prevent them from going over digital 100% (since they had to be increased in gain). Pre-encoded programs can be included in broadcasts if they have had the value of dialnorm correctly set, and the receiver will then conform the level.

The second (and generally preferred) method is to let all programming enter the encoder at full level, and correct for differing levels by adjusting the encoded value of dialnorm to be correct for each program. In this case, the conforming to a common level is done at the receiver. This method will become more practical as computer remote control of the encoding equipment becomes commonplace. The data base for each audio program to be encoded would include (along with items such as number of channels, language, etc.) the dialogue level. The master control computer would then communicate the value of dialogue level to the audio encoder which would then place the appropriate value in the bit stream.

In the case where a complete audio program is formed from the combination of a main and an associated service, each of the two services being combined will have a value of dialnorm, and the values may not be identical. In this case, the value of dialnorm in each bit stream should be used to alter the level of the audio decoded from that bit stream, prior to the mixing process which combines the audio from the two bit streams to form the complete audio program.

6.5.2 Dynamic range compression

It is common practice for high quality programming to be produced with wide dynamic range audio, suitable for the highest quality audio reproduction environment. Broadcasters, serving a wide audience, typically process audio in order to reduce its dynamic range. The processed audio is more suitable for the majority of the audience which does not have an audio reproduction environment which matches that of the original audio production studio. In the case of NTSC, all viewers receive the same audio with the same dynamic range, and it is impossible for any viewer to enjoy the original wide dynamic range audio production.

The audio coding system provides an embedded dynamic range control system which allows a common encoded bit stream to deliver programming with a dynamic range appropriate for each individual listener. A dynamic range control value (dynrng) is provided in each audio block (every 5 ms). These values are used by the audio decoder in order to alter the level of the reproduced audio for each audio block. Level variations of up to ± 24 dB may be indicated. The values of dynrng are generated in order to provide a subjectively pleasing but restricted dynamic range. The unaffected level is dialogue level. For sounds louder than dialogue, values of dynrng will indicated gain reduction. For sounds quieter than dialogue, values of dynrng will indicate a gain increase. The broadcaster is in control of the values of dynrng, and can supply values which generated the amount of compression which the broadcaster finds appropriate. The use of dialogue level as the unaffected level further improves loudness uniformity.

By default, the values of dynrng will be used by the audio decoder. The receiver will thus reproduce audio with a reduced dynamic range, as intended by the broadcaster. The receiver may also offer the viewer the option to scale the value of dynrng in order to reduce the effect of the dynamic range compression which was introduced by the broadcaster. In the limiting case, if the value of dynrng is scaled to zero, then the audio will be reproduced with its full original dynamic range. The optional scaling of dynrng can be done differently for values indicating gain reduction (which reduces the levels of loud sounds) and for values indicating gain increases (which makes quiet sounds louder). Thus the viewer may be given independent control of the amount of compression applied to loud and quiet sounds. Therefore, while the broadcaster may introduce dynamic range compression to suit the needs of most of the audience, individual listeners may have the option to choose to enjoy the audio program with more or all of its original dynamic range intact.

The dynamic range control words may be generated by the AC-3 encoder. They may also be generated by a processor located before or after the encoder. If the dynamic range processor is located prior to the encoder, there is a path to convey the dynamic range control words from the processor to the encoder, or to a bit stream processor, so that the control words may be inserted into the bit stream. If the dynamic range processor is located after the encoder, it can act upon an encoded stream and directly insert the control words without altering the encoded audio. In general, encoded bit streams may have dynamic range control words inserted or modified without affecting the encoded audio.

When it is necessary to alter subjectively the dynamic range of audio programs, the method built into the audio coding subsystem should be used. The system should provide a transparent pathway, from the audio program produced in the audio post production studio, into the home. Signal processing devices such as compressors or limiters which alter the audio signal should not be inserted into the audio signal chain. Use of the dynamic range control system embedded within the audio coding system allows the broadcaster or program provider to appropriately limit the delivered audio dynamic range without actually affecting the audio signal itself. The original audio is delivered intact and is accessible to those listeners who wish to enjoy it.

In the case where a complete audio program is formed from the combination of a main and an associated service, each of the two services being combined may have a dynamic range control signal. In most cases, the dynamic range control signal contained in a particular bit stream applies to the audio channels coded in that bit stream. There are two exceptions: a single channel visually impaired (VI) associated service containing only a narrative describing the picture content, and a voiceover (VO) associated service. In these two cases, the dynamic range control signal in the associated service is used by the decoder to control the audio level of the main audio service. This allows the provider of the VI or VO service the ability to alter the level of the main audio service in order to make the VI or VO services intelligible. In these cases the main audio service level is controlled by both the control signal in the main service and the control signal in the associated service.

6.6 Main, associated, and multi-lingual services

6.6.1 Overview

An AC-3 elementary stream contains the encoded representation of a single audio service. Multiple audio services are provided by multiple elementary streams. Each elementary stream is conveyed by the transport multiplex with a unique PID. There are a number of audio service types which may (individually) be coded into each elementary stream. Each elementary stream is tagged as to its service type using the bsmod bit field. There are two types of *main service* and six types of *associated service*. Each associated service may be tagged (in the AC-3 audio descriptor in the transport PSI data) as being associated with one or more main audio services. Each AC-3 elementary stream may also be tagged with a language code.

This Section describes each type of service and gives usage guidelines. In general, a complete audio program (what is presented to the listener over the set of loudspeakers) may consist of a main audio service, or a main audio service combined with one associated audio service. The capability to simultaneously decode one main service and one associated service is required in order to form a complete audio program in certain service combinations described in this Section.

6.6.2 Summary of service types

The service types which correspond to each value of bsmod are defined in the Digital Audio Compression (AC-3) Standard and in Annex B of the Digital Television Standard. The information is reproduced in Table 6.1 and the following paragraphs briefly describe the meaning of these service types.

Table 6.1 Table of Service Types

bsmod	Type of service
000 (0)	Main audio service: complete main (CM)
001 (1)	Main audio service: music and effects (ME)
010 (2)	Associated service: visually impaired (VI)
011 (3)	Associated service: hearing impaired (HI)
100 (4)	Associated service: dialogue (D)
101 (5)	Associated service: commentary (C)
110 (6)	Associated service: emergency (E)
111 (7)	Associated service: voice-over (VO)

6.6.2.1 Complete main audio service (CM)

This is the normal mode of operation. All elements of a complete audio program are present. The audio program may be any number of channels from 1 to 5.1.

6.6.2.2 Main audio service, music and effects (ME)

All elements of an audio program are present except for dialogue. This audio program may contain from 1 to 5.1 channels. Dialogue from a D associated service must be simultaneously decoded and added to form a complete program.

6.6.2.3 Associated service: visually impaired (VI)

This is typically a single channel service, intended to convey a narrative description of the picture content for use by the visually impaired, that is to be decoded along with the main audio service. The VI service may be a complete mix of all program elements, in which case it may use any number of channels (up to 5.1).

6.6.2.4 Associated service: hearing impaired (HI)

This is typically a single channel service, intended to convey dialogue which has been processed for increased intelligibility for the hearing impaired, that is to be decoded along with the main audio service. The HI service may be a complete mix of all program elements, in which case it may use any number of channels (up to 5.1).

6.6.2.5 Associated service: dialogue (D)

This service conveys dialogue intended to be mixed into a main audio service (ME) which does not contain dialogue.

6.6.2.6 Associated service: commentary (C)

This service conveys a single channel of commentary. This commentary channel differs from a dialogue service, in that it contains optional instead of necessary program content.

6.6.2.7 Associated service: emergency message (E)

This is a single channel service, which is given priority in reproduction. If this service type appears in the transport multiplex, it is routed to the audio decoder. If the audio decoder receives this service type, it will decode and reproduce the E channel while muting the main service.

6.6.2.8 Associated service: voice-over (VO)

This is a single channel service intended to be decoded and added into the center loudspeaker channel.

6.6.3 Multi-lingual services

Each audio bit stream may be in any language. In order to provide audio services in multiple languages a number of main audio services may be provided, each in a different language. This is the (artistically) preferred method, because it allows unrestricted placement of dialogue along with the dialogue reverberation. The disadvantage of this method is that as much as 384 kbps is needed to provide a full 5.1 channel service for each language. One way to reduce the required bit-rate is to reduce the number of audio channels provided for languages with a limited audience. For instance, alternate language versions could be provided in 2-channel stereo with a bit-rate of 128 kbps. Or, a mono version can be supplied at a bit-rate of approximately 64-96 kbps.

Another way to offer service in multiple languages is to provide a main multi-channel audio service (ME) which does not contain dialogue. Multiple single channel dialogue associated services (D) can then be provided, each at a bit-rate of approximately 64-96 kbps. Formation of a complete audio program requires that the appropriate language D service be simultaneously decoded and mixed into the ME service. This method allows a large number of languages to be efficiently provided, but at the expense of artistic limitations. The single channel of dialogue would be mixed into the center reproduction channel, and could not be panned. Also, reverberation would be confined to the center channel, which is not optimum. Nevertheless, for some types of programming (sports, etc.) this method is very attractive due to the savings in bit rate it offers.

Stereo (two channel) service without artistic limitation can be provided in multiple languages with added efficiency by transmitting a stereo ME main service along with stereo D services. The D and appropriate language ME services are simply combined in the receiver into a complete stereo program. Dialogue may be panned, and reverberation may be placed included in both channels. A stereo ME service can be sent with high quality at 192 kbps, while the stereo D services (voice only) can make use of lower bit-rates, such as 128 or 96 kbps per language.

Note that during those times when dialogue is not present, the D services can be momentarily removed, and their data capacity used for other purposes.

6.6.4 Detailed description of service types

6.6.4.1 CM — complete main audio service

The CM type of main audio service contains a complete audio program (complete with dialogue, music, and effects). This is the type of audio service normally provided. The CM service may contain from 1 to 5.1 audio channels. The CM service may be further enhanced by means of the VI, HI, C, E, or VO associated services described below. Audio in multiple languages may be provided by supplying multiple CM services, each in a different language.

6.6.4.2 ME — main audio service, music and effects

The ME type of main audio service contains the music and effects of an audio program, but not the dialogue for the program. The ME service may contain from 1 to 5.1 audio channels. The primary program dialogue is missing and (if any exists) is supplied by decoding simultaneously a D service. Multiple D services in different languages may be associated with a single ME service.

6.6.4.3 VI — visually impaired

The VI associated service typically contains a narrative description of the visual program content. In this case, the VI service is a single audio channel. Simultaneous decoding of the VI service and the main audio service allows the visually impaired user to enjoy the main multi-channel audio program, as well as to follow the on-screen activity. This allows the VI service to be mixed into one of the main reproduction channels (the choice of channel may be left to the listener) or to be provided as a separate output (which, for instance, might be delivered to the VI user via open-air headphones).

The dynamic range control signal in the VI service is intended to be used by the audio decoder to modify the level of the main audio program. Thus the level of the main audio service will be under the control of the VI service provider, and the provider may signal the decoder (by altering the dynamic range control words embedded in the VI audio elementary stream) to reduce the level of the main audio service by up to 24 dB in order to assure that the narrative description is intelligible.

Besides providing the VI service as a single narrative channel, the VI service may be provided as a complete program mix containing music, effects, dialogue, and the narration. In this case, the service may be coded using any number of channels (up to 5.1). The fact that the service is a complete mix is indicated in the AC-3 descriptor (see Section 5.7.2.1 in Annex C of the Digital Television Standard).

6.6.4.4 HI — hearing impaired

The HI associated service typically contains only a single channel of dialogue and is intended for use by those whose hearing impairments make it difficult to understand the dialogue in the presence of music and sound effects. The dialogue may be processed for increased intelligibility by the hearing impaired. The hearing impaired listener may wish to listen to a mixture of the single channel HI dialogue track and the main program audio. Simultaneous decoding of the HI service along with the CM service allows the HI listener to adjust the mixture to control the emphasis on dialogue over music and effects. The HI channel would typically be mixed into the center channel. An alternative would be to deliver the HI signal to a discrete output (which, for instance, might feed a set of open-air headphones worn only by the HI listener.)

Besides providing the HI service as a single narrative channel, the HI service may be provided as a complete program mix containing music, effects, and dialogue with enhanced intelligibility. In this case, the service may be coded using any number of channels (up to 5.1). The fact that the service is a complete mix shall be indicated in the AC-3 descriptor (see Section 5.7.2.1 in Annex C of the Digital Television Standard).

6.6.4.5 D — dialogue

The dialogue associated service is employed when it is desired to most efficiently offer multi-channel audio in several languages simultaneously, and the program material is such that the restrictions (no panning, no multi-channel reverberation) of a single dialogue channel may be tolerated. When the D service is used, the main service is of type ME (music and effects). In the case that the D service contains a single channel, simultaneously decoding the ME service along with the selected D service allows a complete audio program to be formed by mixing the D channel into the center channel of the ME service. Typically, when the main audio service is of type ME, there will be several different language D services available. The transport demultiplexer may be designed to select the appropriate D service to deliver to the audio decoder based on the listener's language preference (which would typically be stored in memory in the receiver). Or, the listener may explicitly instruct the receiver to select a particular language track, overriding the default selection.

If the ME main audio service contains more than two audio channels, the D service will be monophonic (1/0 mode). If the main audio service contains two channels, the D service may contain two channels (2/0 mode). In this case, a complete audio program is formed by simultaneously decoding the D service and the ME service, mixing the left channel of the ME service with the left channel of the D service, and mixing the right channel of the ME service with the right channel of the D service. The result will be a two channel stereo signal containing music, effects, and dialogue.

6.6.4.6 C — commentary

The commentary associated service is similar to the D service, except that instead of conveying primary program dialogue, the C service conveys optional program commentary. When C service(s) are provided, the receiver may notify the listener of their presence. The listener should be able to call up information (probably on-screen) about the various available C services, and optionally request one of them to be selected for decoding along with the main service. The C service may be added to any loudspeaker channel (the listener may be given this control). Typical uses for the C service might be optional added commentary during a sporting event, or different levels (novice, intermediate, advanced) of commentary available to accompany documentary or educational programming.

6.6.4.7 E — emergency

The E associated service is intended to allow the insertion of emergency announcements. The normal audio services do not necessarily have to be replaced in order for the emergency message to get through. The transport demultiplexer shall give first priority to this type of audio service. Whenever an E service is present, it is delivered to the audio decoder by the transport subsystem. When the audio decoder receives an E type associated service, it stops reproducing any main service being received and only reproduces the E service. The E service may also be used for non-emergency applications. It may be used whenever the broadcaster wishes to force all decoders to quit reproducing the main audio program and substitute a higher priority single channel.

6.6.4.8 VO — voice-over

It is possible to use the E service for announcements, but the use of the E service leads to a complete substitution of the voice-over for the main program audio. The voice-over associated service is similar to the E service, except that it is intended to be reproduced along with the main service. The systems demultiplexer shall give second priority to this type of associated service (second only to an E service). The VO service is intended to be simultaneously decoded and mixed into the center channel of the main audio service which is being decoded. The dynamic range control signal in the VO service is intended to be used by the audio decoder to modify the level of the main audio program. Thus the level of the main audio service will be under the control of the broadcaster, and the broadcaster may signal the decoder (by altering the dynamic range control words embedded in the VO audio bit stream) to reduce the level of the main audio service by up to 24 dB during the voice over. The VO service allows typical voice-overs to be added to an already encoded audio bit stream, without requiring the audio to be decoded back to baseband and then re-encoded. However, space must be available within the transport multiplex to make room for the insertion of the VO service.

6.7 Audio bit rates

6.7.1 Typical audio bit rates

The following information provides a general guideline as to the audio bit rates which are expected to be most useful. For main services, the use of the LFE channel is optional and will not affect the indicated data rates.

Table 6.2 Typical Audio Bit Rate

Type of service	Number of channels	Typical bit rates
CM, ME	5	320-384 kbps
CM, ME	4	256-384 kbps
CM, ME	3	192-320 kbps
CM, ME	2	128-256 kbps
VI	1	48-128 kbps
HI	1	48-96 kbps
D	1	64-128 kbps
D	2	96-192 kbps
C	1	32-128 kbps
E	1	32-128 kbps
VO	1	64-128 kbps

6.7.2 Audio bit rate limitations

The audio decoder input buffer size (and thus part of the decoder cost) is determined by the maximum bit rate which must be decoded. The syntax of the AC-3 standard supports bit rates ranging from a minimum of 32 kbps up to a maximum of 640 kbps per individual elementary bit stream. The bit rate utilized in the digital television system is restricted in order to reduce the size of the input buffer in the audio decoder, and thus the receiver cost. Receivers will support the decoding of a main audio service at a bit rate up to and including 384 kbps. Transmissions may contain main audio services encoded at a bit rate up to and including 384 kbps. Transmissions may contain single channel associated audio services (intended to be simultaneously decoded along with a main service) encoded at a bit rate up to and including 128 kbps. Transmissions may contain dual channel associated services (intended to be simultaneously decoded along with a main service) encoded at a bit rate up to and including 192 kbps. Transmissions have a further limitation that the combined bit rate of a main and an associated service which are intended to be simultaneously decoded is less than or equal to 512 kbps.